DOCUMENT RESUME

ED 129 229 IR 004 005

TITLE Subject Access Project; First Quarterly Progress Report.
INSTITUTION Syracuse Univ., N.Y. School of Information Studies.
PUB DATE Sep 76
NOTE 16p.; Includes Occasional Newsletter no. 1, August 1976

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Cataloging; *Experimental Programs; Humanities; *Indexing; Information Retrieval; Library Research; Social Sciences; *Subject Index Terms; University Libraries
IDENTIFIERS Machine Readable Cataloging; *MARC; *Subject Access

ABSTRACT

An attempt is being made to improve subject access to monographs by augmenting MARC (Machine Readable Cataloging) records. Working with a sample of books drawn from the collections of the University of Toronto and comprising a number of subject areas in the humanities and social sciences, the project plans to enlarge the MARC description by using a set of decision rules for selecting words and phrases found in the index and/or the table of contents. This first quarterly report for the period June to August 1976 describes the sample of selected monographs, summarizes the project budget to date, and projects activities for the future. (EMH)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED129229

SYRACUSE UNIVERSITY

SCHOOL OF INFORMATION STUDIES

SUBJECT ACCESS PROJECT

FIRST QUARTERLY PROGRESS REPORT *

PROJECT STAFF

Pauline Atherton - Project Director
B.K.L. Genova - Assistant Project Director
Bette Brindle - Research Associate
Barbara Settel - Research Assistant
Cheryl Willson - Secretary

and members of the University of Toronto, Robarts Library staff who helped on site during the sampling and photocopying steps.

Time Table for Key Events in Project - First Quarter

1. Sample selection at University of Toronto to represent several subject categories in depth. Photocopy sample monographs' table of contents and indexes.

2. Descriptive report on sample collection. For example:
   Number of titles per subject categories selected
   Number of useful contents pages
   Number of subject indexes found (pages of index/book), etc.

* Due to funding delay, the first Quarterly Progress Report covers the period from June 15 through August 31, 1976.

IR 004 005

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SAMPLE OF UNIVERSITY OF TORONTO LIBRARY MACHINE-READABLE CATALOG
(Acquisitions since 1966)

The emphasis of our project is on experimentation with improved subject access to books in the humanities and social sciences. As a part of this effort, our goal in sampling post-1966 acquisitions at the University of Toronto is the ability to outline reliably the indexing profile of monographs representing fields of knowledge we would work with. We did not aim at a sample representative of the Library's entire holdings in the social sciences and humanities. Rather, we opted for in-depth study through a representative sample of eight disciplines divided equally among the above two broad areas. Eligible items for study in our project were monographs in English only. We used the UTLAS Conversion Project listings as the sampling frame, which was initially sorted to exclude foreign language materials and serials, as coded by the University of Toronto Library. We then drew a systematic random sample of 1592 items, as seen in Table 1. We believe this selection presents a good balance between the areas of knowledge concerning us, with an eye on potential user demand in a typical academic setting.

In addition, we decided to study two selections in their entirety; i.e., exhaustive listings of the University of Toronto, Robarts Library holdings in Post-Confederation Ontario History (300 items) and Urban Planning (523 items). These two selections will allow thorough searching for real requests posed to the Robarts Library reference staff.

Taken together, we will work with augmented records of 2415 monographs. The sample and the two selections will be kept separate for purposes of analysis.

TABLE 1*

Sample Composition, n = 1592** and Complete Selections (n = 823). (Grand Total: 2415 items)

Humanities	# of items
1. Philosophy: BC Logic BH Aesthetics BJ Ethics	152
2. History: DE Graeco-Roman World DF Ancient Greece DG Ancient Italy Rome to 476	139
3. Arts: NB Sculpture NE Engraving NK Art Applied to Industry Decoration & Ornament	201
4. Literature: PN 1560-3300 Drama	320
TOTAL Humanities	812

Social Sciences	# of items
5. Psychology: BF 1-990	346
6. Anthropology: GN 1-696	144
7. Public Finance: HJ	136
8. Sociology: HM 1-221	154
TOTAL Social Sciences	780
Complete Selection for Two Fields	
1. Urban Planning. Urban Redevelopment: HT 166-177	523
2. Post-Confederation Ontario History: F 5520-5547	300

* This is updated version of table presented in August "Occasional Newsletter."
** Total estimated population of English language, non-serial items N=16,910, with 8,601 (or 51%) in the Humanities, and 8,310 (or 49%) in the Social Science disciplines.

Our first "Occasional Newsletter" (August 1976) presented the structure of our sample drawn using the UTLAS Conversion Project listings, in preparation for the preliminary analysis. Since then we have continued photocopying indexes, tables of contents, illustration lists, maps, etc. from each monograph, in order to prepare results as follows:

1) actual number of monographs selected

2) number of monographs with an index

3) number of monographs without an index, table of contents, or lists

4) number of monographs that contain a table of contents but no index

5) number of monographs that have no table of contents or index but have some kind of list of illustrations, maps, or tables

6) Attrition rate - number of missing items and deletes

7) Profiles of the contents and index material for each class of monographs

Our progress report at this time (September 1976) outlines the results we have compiled so far. One hundred twenty-five items have either not been located or fully processed as yet. Incoming data regarding their indexes and tables of contents would change the profiles generated as of now. Some of these 125 items may fall into the sixth category: attrition. To date, 61 items have been deleted from the original sample as duplicates, non-English books, or journal articles erroneously listed as monographs for the UTLAS Conversion Project. Refer to Table 1, the record of our sample.

Table Two below reports the current status of monograph analysis as follows:
Column 1 - number of items located, processed and ready for the next phase.
Column 2 - indicates the remaining items that must be located and processed; if not found, deleted.
Column 3 - number of processed items in Column 1 that have usable table of contents
Column 4 - percent of the processed that contain table of contents
Column 5 - number of processed (Column 1) items that have an index
Column 6 - percent of processed (Column 1) items that have an index
Column 7 - the range of the number of pages per index

Included in the appendix are histograms of occurrences of number of pages per index. The mode for all disciplines is below 10 pages of index, a result which makes our work (next quarter) very manageable.

TABLE 2

Sample	# of Items (1)	Not Processed (2)	Table of Contents #(3)	Table of Contents (4)	Index #(5)	Index (6)	Range of # of Index Pages (7)*
1. Philosophy BC Logic BH Aesthetics BJ Ethics	143	9	134	94%	80	56%	1-74
2. History DE Graeco-Roman World DF Ancient Greece DG Ancient Italy Rome to 476	126	13	120	95	99	79	1-141
3. Arts NB Sculpture NE Engraving NK Art Applied to Industry Decoration & Ornament	185	16	133	72	89	48	2-94
4. Literature PN 1560-3300	296	24	255	86	154	52	1-35
5. Psychology BF 1-990	318	28	309	97	250	79	1-117
6. Anthropology GN 1-696	138	6	130	94	93	67	1-42
7. Public Finance HJ	127	9	107	84	70	55	1-76
8. Sociology HM 1-221	148	6	141	95	102	69	1-27
Sample Totals	1481	111	1329	90%	937	63%	1-141
Complete Collection Listings							
1. Urban Planning. Urban Redevelopment HT 166-177	516	7	467	91	146	28	1-43
2. Post Confederation Ontario History F 5520-5547	293	7	190	65	88	30	1-89
Complete Collection Total	809	14	657	81%	234	29%	1-89
Grand Total	2290	125	1986	87%	1171	51%	1-141

* See Histograms in Appendix

ANALYSIS OF THE CHARACTERISTICS OF INDEXES AND CONTENTS PAGES IN BOOKS

During the past two-three weeks we have been engaged in the process of observing and analyzing the composition of indexes and tables of contents from the monographs in all fields sampled. An overview of the variety of syntactical and physical relationships is in the draft stage now. At this point, we are not willing to make any statement of generalities describing or defining indexes and their elements. These are non-standardized, free-text, indexes and contents lists.

The indexes and tables of contents are being approached in several ways. Comparisons may be made:

(a) within one subject field,
(b) by type of index (subject or subject & name) without regard to discipline, or
(c) between the table of contents and the index entries of one book.

In quantitative terms, we are looking at the percent of "names" versus "subject" entries, the extent to which page ranges after an index entry occur, the possible relationship between subdivided index entries and the length and occurrence of page ranges, the approximate number of entries per chapter, per page, or for an entire book.

Qualitative assessments are being focused on semantic relationships between words and phrases. In this respect, most notable are the number of entries which can only be properly understood within the context of the entire index or book itself. Words like <u>the</u>, <u>then</u>, which are automatically assigned to stop word lists for computerized subject searching cannot be easily discarded in searching index entries in books on logic where entries like "if....then," "law, the moral" appear.

We will continue these studies and begin comparisons across subject fields, all in preparation for our choice of selection rules.

INFORMAL FINANCIAL REPORT

SALARIES		Budget Total	Expended as of 8/31/76
Project Director		14,458	6,527.00
Assoc. Project Director (University salary increase as of 7/1/76)		4,950	5,206.00
Grad Res. Assts. 50% AY		4,940	
100% Summer		4,060	2,925.[illegible]5
Secretary		6,500	926.29
Keypunchers (Hourly)		4,500	
Survey Assts. (Hourly)		4,000	453.75
Programmer (Hourly)		100	
	Subtotal	43,508	16,038.79
FRINGE BENEFITS @ 23.5%		10,225	3,769.12
	Subtotal	53,733	19,807.91
OTHER DIRECT COSTS			
Travel		1,995	1,285.63
Duplicating		3,000	909.15
Supplies (Office)		1,500	113.37
Equipment Rental			
EDP		1,850	
Office		2,542	290.00
SERVICES			
EDP		1,[illegible]0	121.48
SDC Sub-Contract		7,955	
Processing		2,170	20.00
	Total	76,615	22,547.54

This represents our best estimate of expenditures to date. It is different from the official report from Syracuse University because of the normal delays in accounting procedures in such a large institution.

Expected Events Somewhat Delayed

1. Arrival of machine-readable tape of items in sample from UTLAS was September 15.

2. Computer terminal. When it arrives, we will perform some searches in LIBCON to assess state-of-art for subject access to books on present on-line retrieval services.

3. Concurrently, we hope to develop separate access, on-line, to a computer tape from BRODART which has enriched subject headings for 800 books from DC 300-389 sections of the Public Library Catalog.

Priority Tasks in Next Quarter

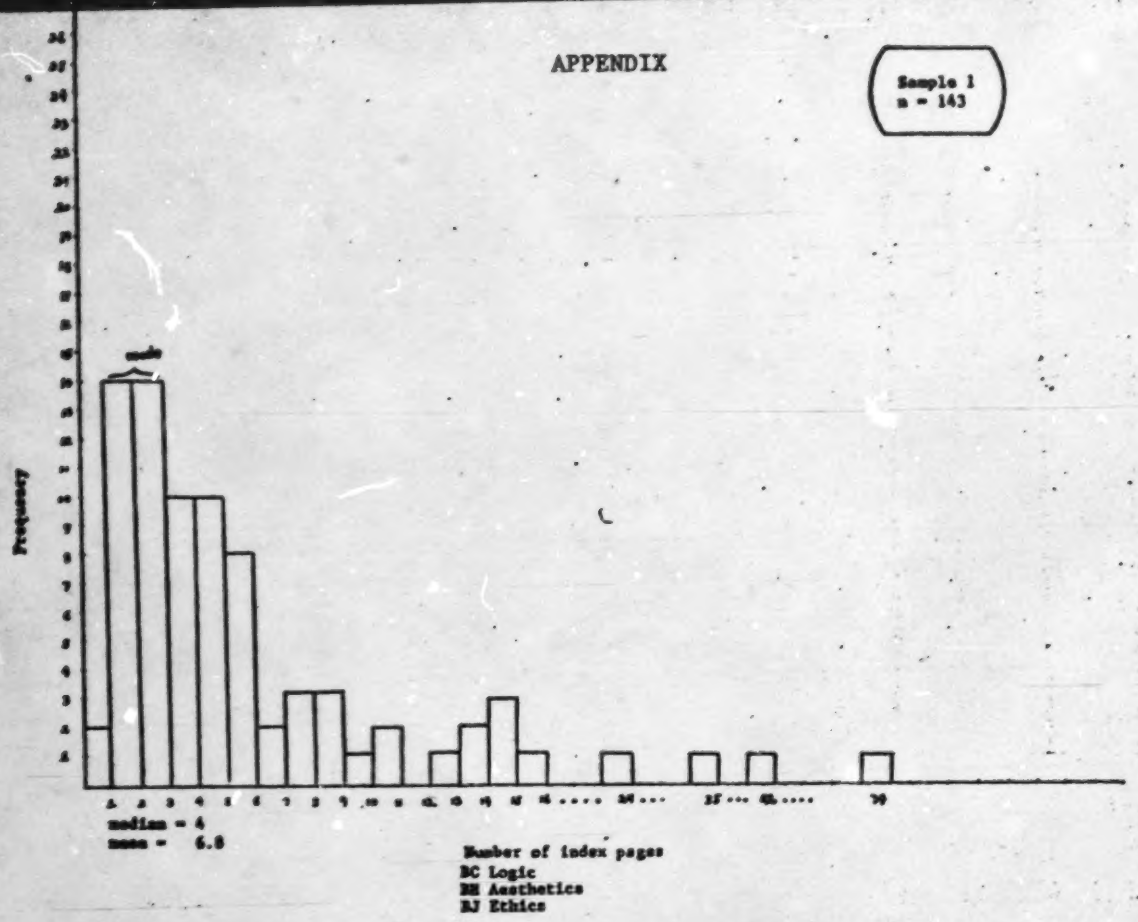
1. Reports on characteristics of subjcet indexes and contents pages followed by development of definitive selection rules.

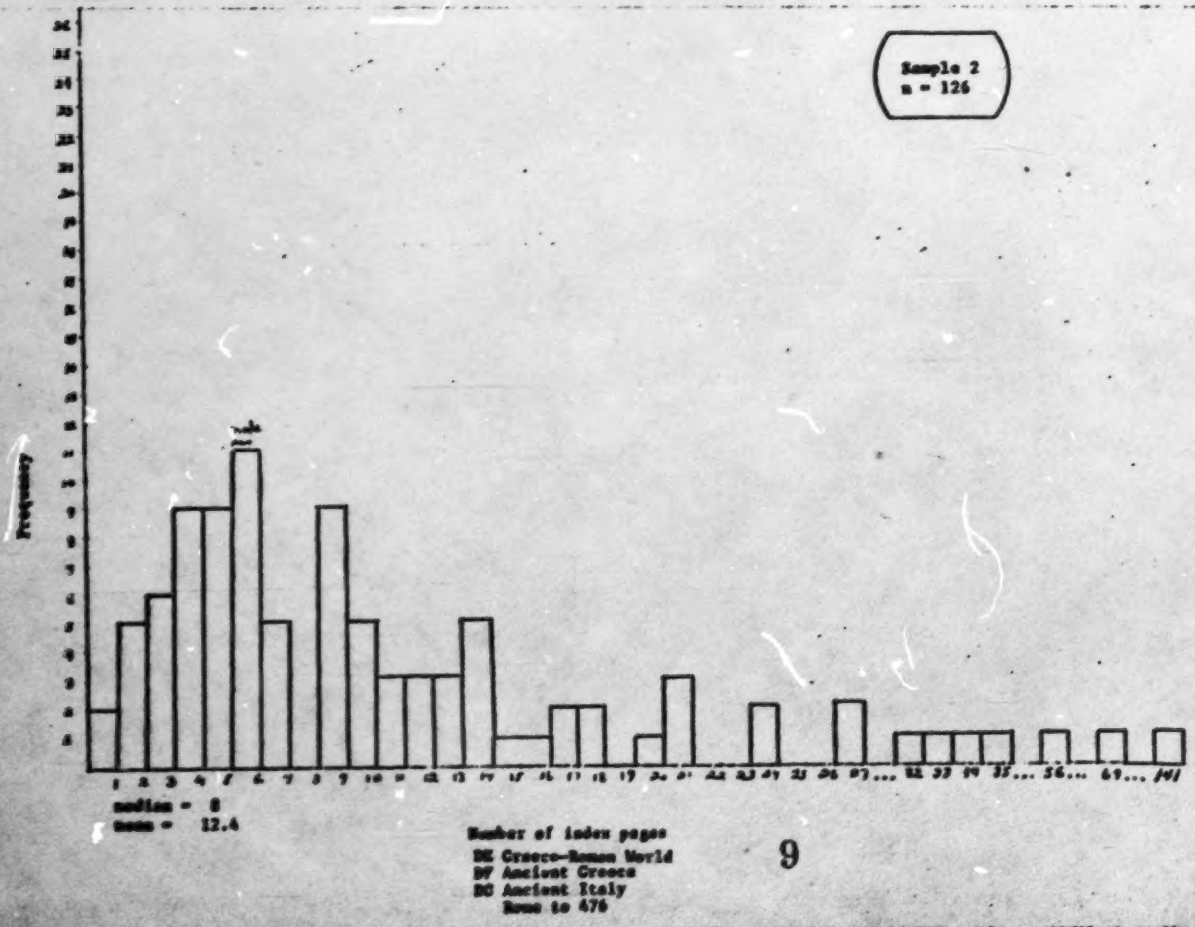
2. Discussion with library users and reference librarians on optimum access and output format options for on-line searching.

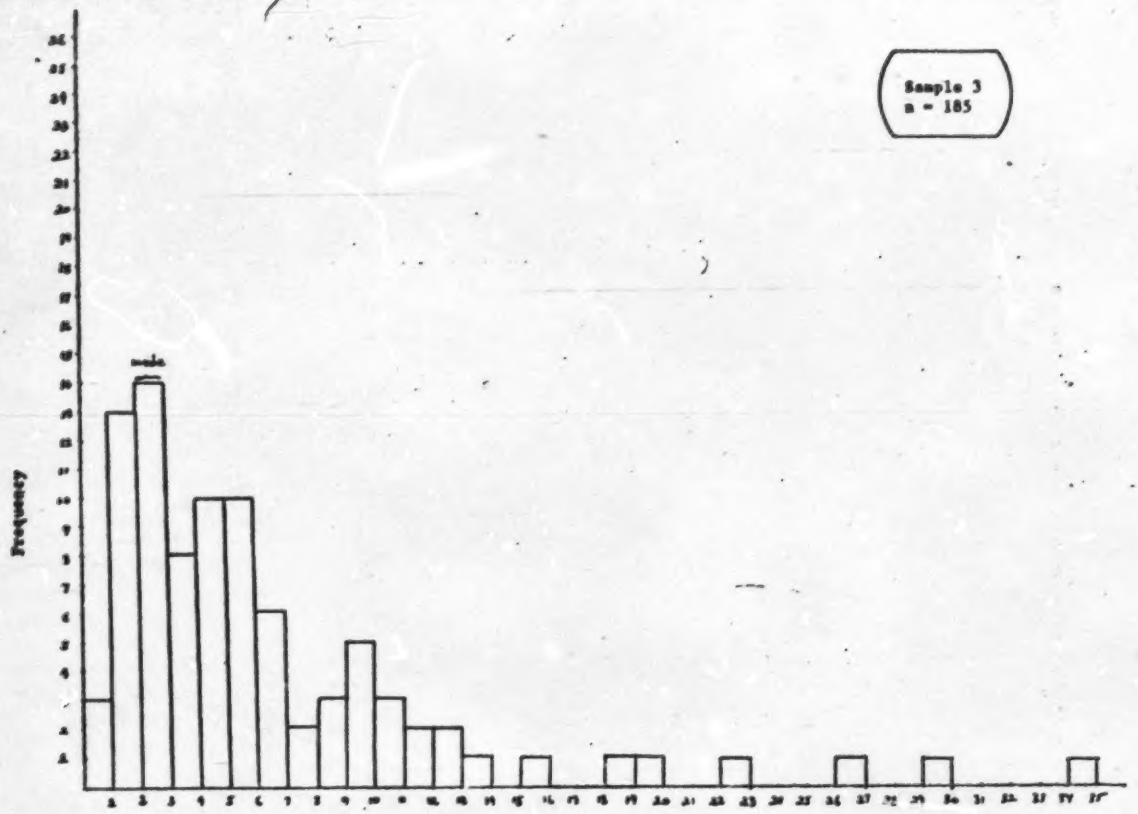
3. Preliminary cost analysis for data input.

4. Preliminary preparation of machine-readable data for SDC/ORBIT system.

APPENDIX

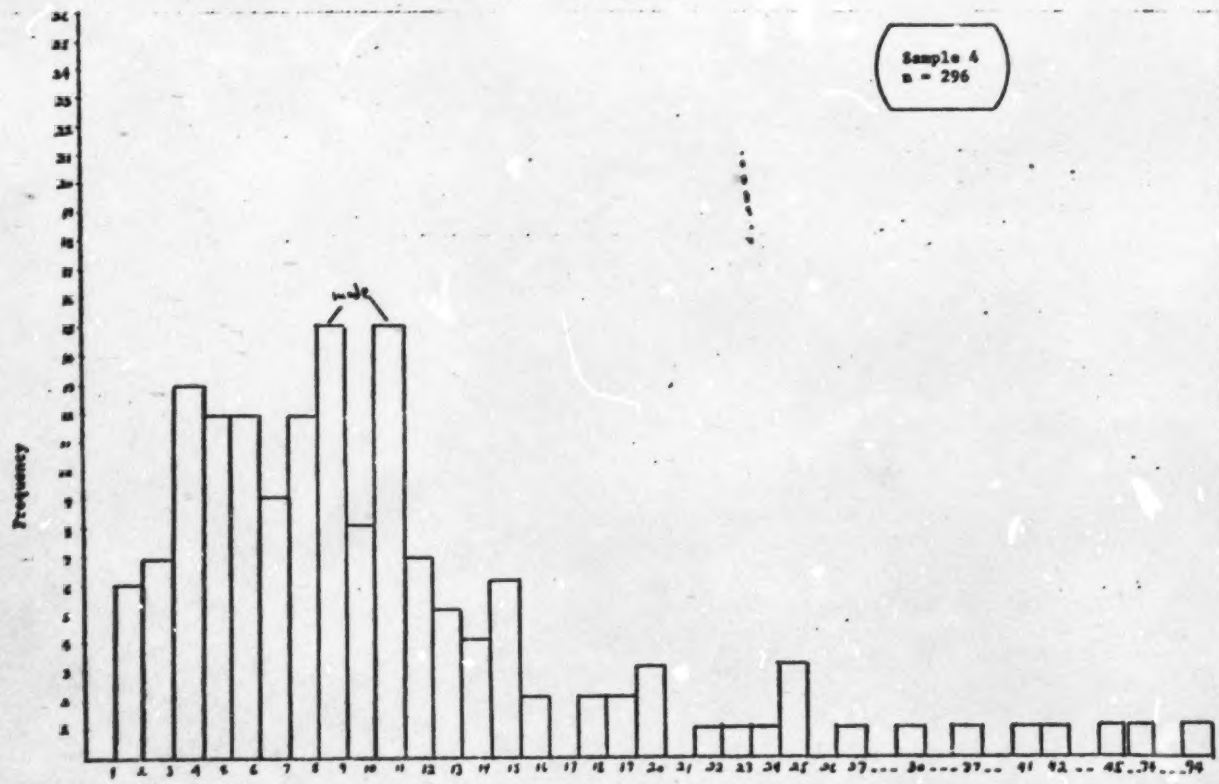

Sample 1
n = 143
Frequency
mode
median = 4
mean = 6.8
Number of index pages
BC Logic
BH Aesthetics
BJ Ethics


Sample 2
n = 126
Frequency
mode
median = 8
mean = 12.4
Number of index pages
DE Graeco-Roman World
DF Ancient Greece
DG Ancient Italy
Rome to 476



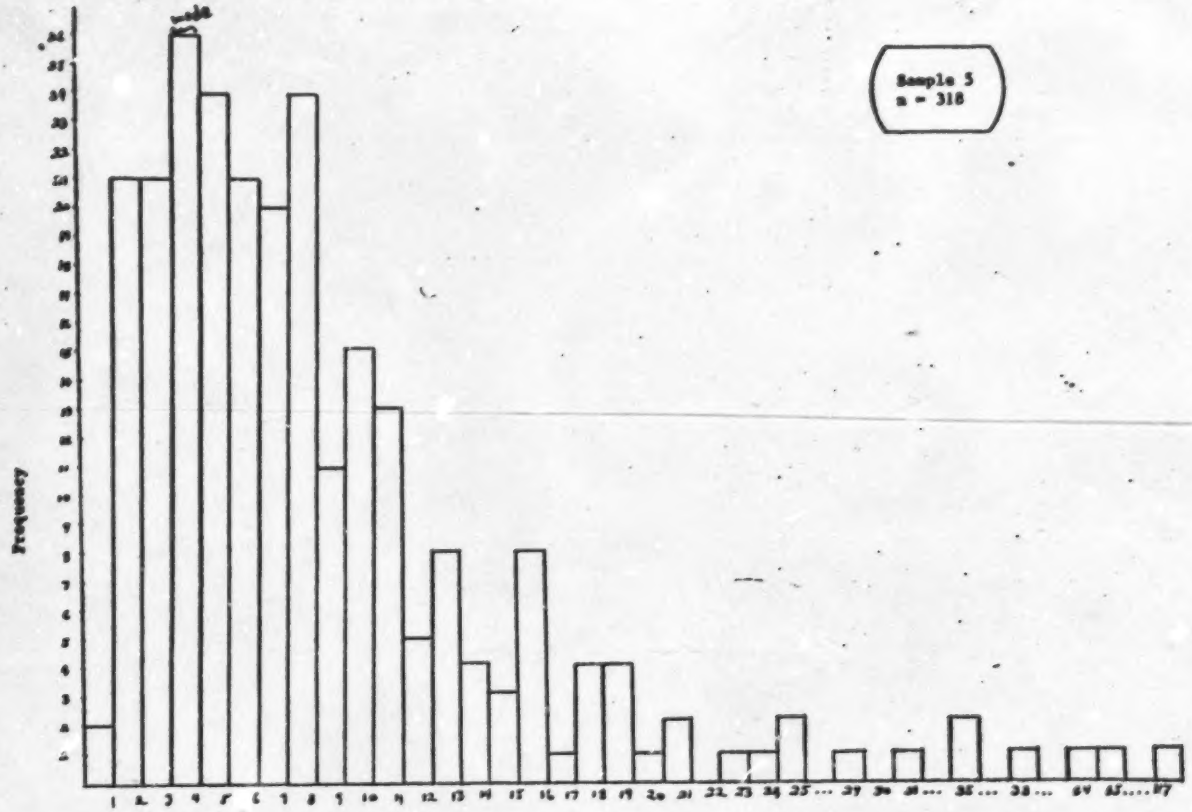
median = 5
mean = 7.2

Number of index pages
NB Sculpture
NE Engraving
NK Art Applied to Industry
Decoration & Ornament



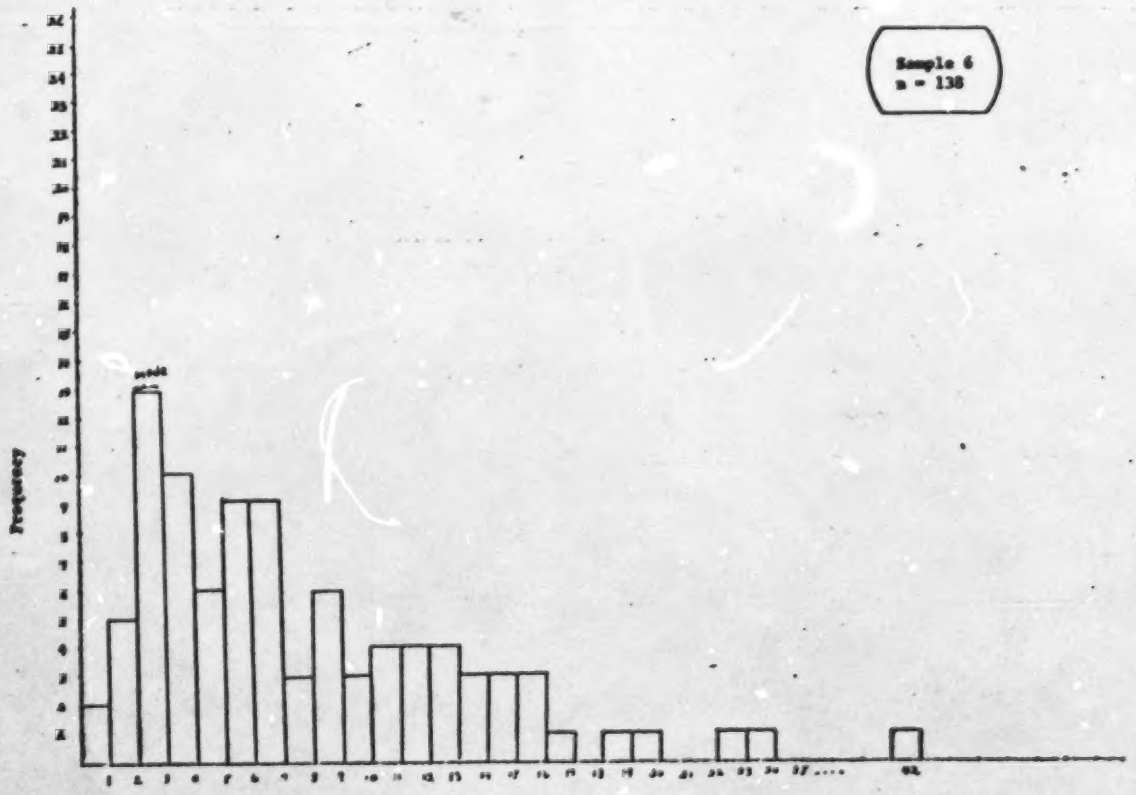
median = 9
mean = 12.2

Number of index pages
PN Drama



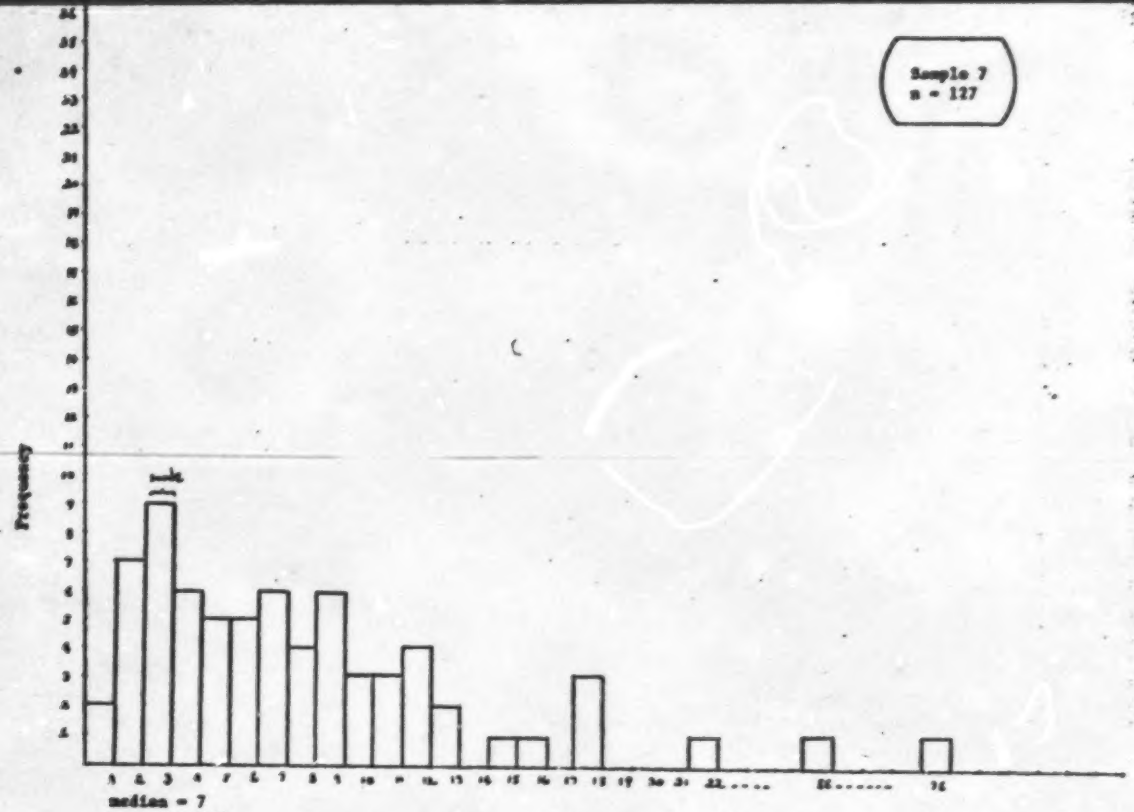
median = 6
mean = 9.2

Number of index pages
BF Psychology

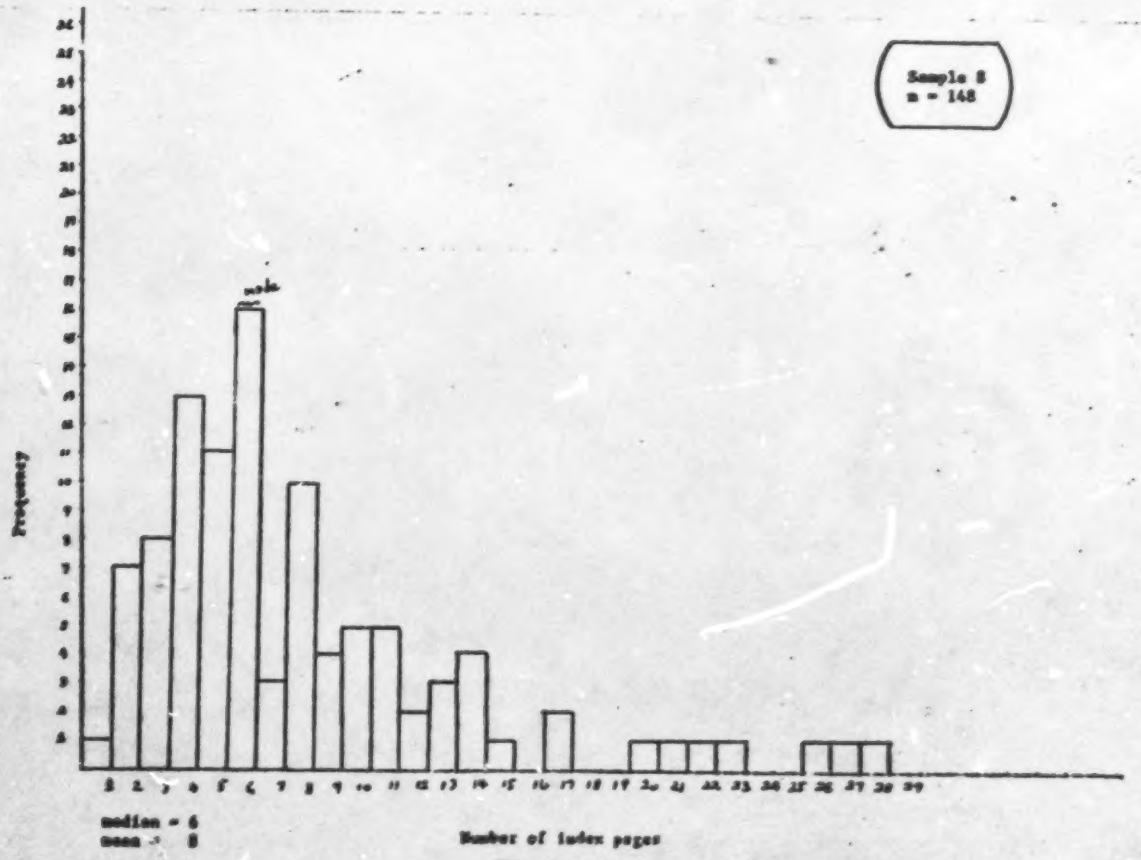


median = 7
mean = 8.3

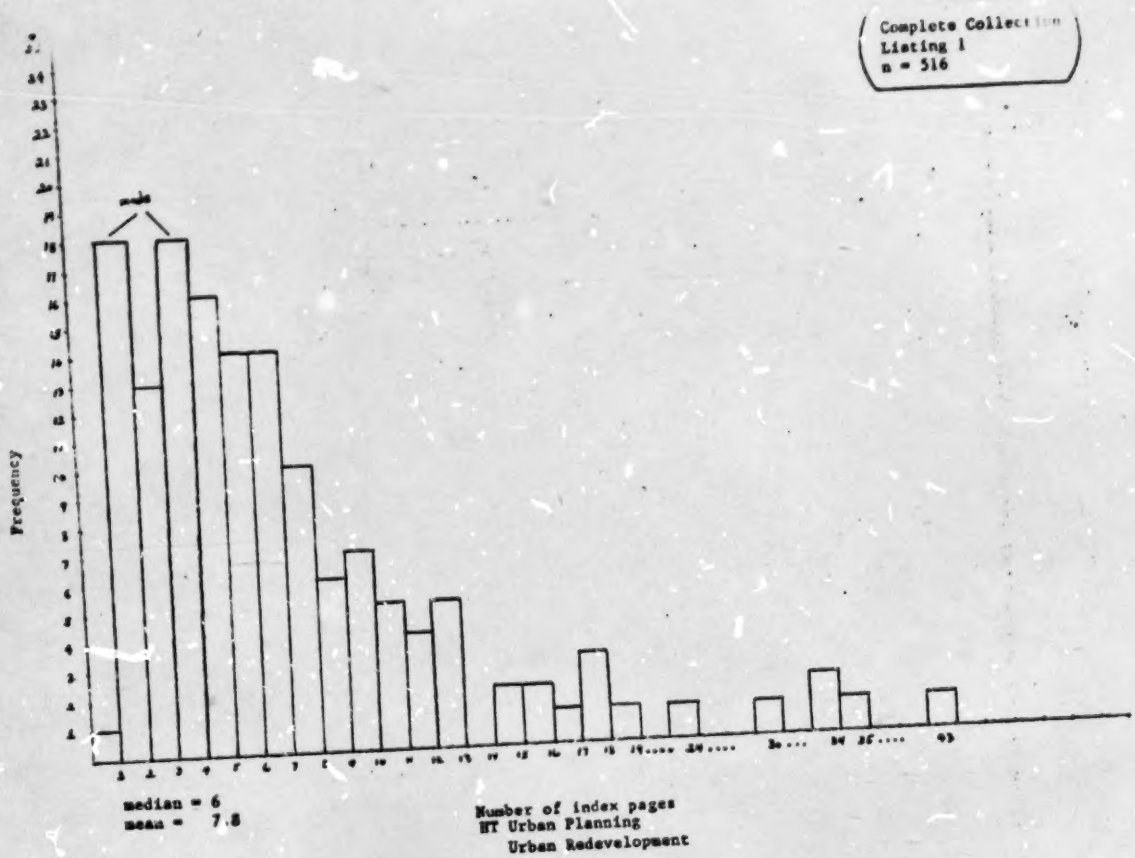
Number of index pages
GN Anthropology

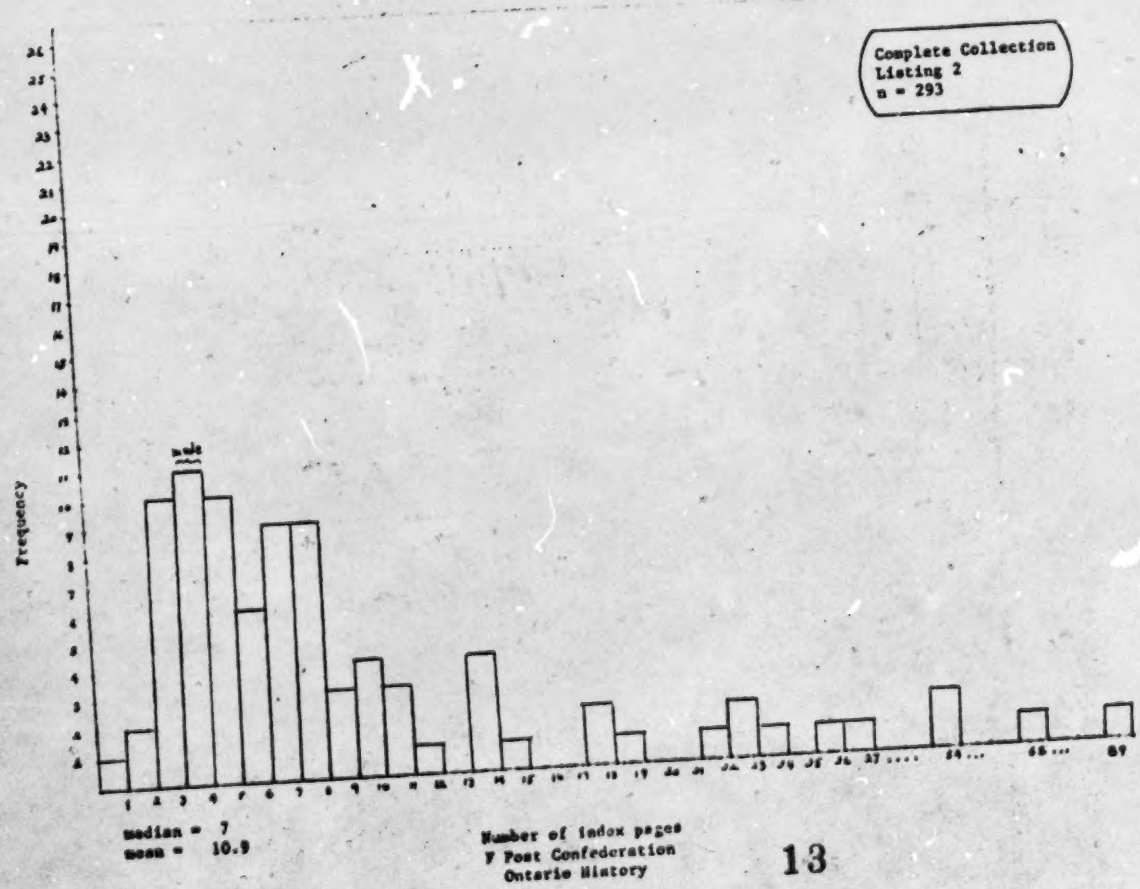


Number of index pages
EJ Public Finance



Number of index pages
AM Sociology


Complete Collection
Listing 1
n = 516
Frequency
median = 6
mean = 7.8
Number of index pages
HT Urban Planning
Urban Redevelopment


Complete Collection
Listing 2
n = 293
Frequency
median = 7
mean = 10.9
Number of index pages
F Post Confederation
Ontario History

Distributed August 15

SYRACUSE UNIVERSITY
SCHOOL OF INFORMATION STUDIES

SUBJECT ACCESS PROJECT

113 EUCLID AVENUE SYRACUSE, NEW YORK 13210 PHONE (315) 423-2911

OCCASIONAL NEWSLETTER

No. 1, August 1976

CLR FUNDS EXPERIMENT AT SYRACUSE UNIVERSITY TO IMPROVE SUBJECT ACCESS TO BOOKS

The Council on Library Resources has awarded $76,615 to the Syracuse University School of Information Studies for an experiment that may result in improved subject access to monographs by augmentation of MARC (Machine Readable Cataloging) records.

Working with a sample of books drawn from the collections of the University of Toronto and comprising a number of subject categories in the humanities and social sciences, the project plan is to enlarge the subject description contained in the MARC record of each book by utilizing a set of selection rules for choosing words and phrases found within the index and/or table of contents. The file of descriptions for the books in the sample will be processed by the System Development Corporation's ORBIT Search Service. On-line computer-based subject searches can then be made by both project staff and others who have access to the Service. The results of the experiment will be analyzed and evaluated to determine the feasibility and utility of performing on-line computer subject searches for monographs with such enriched records.

While the card catalog works well when an author or title is known, it is less efficient for retrieving materials by subject, since only a few broad terms can be used for each item. Several current computer-based abstracting and indexing services have provided far more detailed subject access to journal articles (e.g. Psychological Abstracts, Index Medicus, Engineering Index, etc.) than the MARC records do for books. Little has been done to provide better retrieval capability for the user of monographs. By enriching existing records for monographs, more specific searches in the "free text" mode can be made by searching every word in the subject description.

According to Pauline Atherton, professor in the School of Information Studies and director of the project, if the results are successful, "book indexers and publishers may wish to take greater care in preparing indexes, knowing the use of their efforts for on-line computer-based searching."

PROJECT STAFF

Pauline Atherton - Project Director
B.K.L. Genova - Assistant Project Director
Bette Brindle - Research Associate
Cheryl Willson - Secretary

SUBJECT ACCESS PROJECT
School of Information Studies
Syracuse University
113 Euclid Avenue
Syracuse, New York 13210
(315) 423-2001

SAMPLE OF UNIVERSITY OF TORONTO LIBRARY MACHINE-READABLE CATALOG
(Acquisitions since 1966)

The emphasis of our project is on experimentation with improved subject access to books in the humanities and social sciences. As a part of this effort, our goal in sampling post-1966 acquisitions at the University of Toronto is the ability to outline reliably the indexing profile of monographs representing fields of knowledge we would work with. We did not aim at a sample representative of the Library's entire holdings in the social sciences and humanities. Rather, we opted for in-depth study through a representative sample of eight disciplines divided equally among the above two broad areas. Eligible items for study in our project were monographs in English only. We used the UTLAS Conversion Project listings as the sampling frame, which was initially sorted to exclude foreign language materials and serials, as coded by the University of Toronto Library. We then drew a systematic random sample of 1600 items, as seen in Table 1. We believe this selection presents a good balance between the areas of knowledge concerning us, with an eye on potential user demand in a typical academic setting. In addition, we decided to study two selections in their entirety; i.e., exhaustive listings of the University of Toronto, Robarts Library holdings in Post-Confederation Ontario history (317 items) and Urban Planning (550 items). These two selections will allow thorough searching for real requests posed to the Robarts Library reference staff. Taken together, we will work with augmented records of 2,460 monographs. The sample and the two selections will be kept separate for purposes of analysis.

TABLE 1

Sample composition, n=1600*

Humanities	# of items
1. Philosophy: BC Logic BH Aesthetics BJ Ethics	151
2. History: DE Graeco-Roman World DF Ancient Greece DG Ancient Italy Rome to 476	139
3. Arts: NB Sculpture NE Engraving NK Art Applied to Industry Decoration & Ornament	197
4. Literature: PN 1560-3300 Drama	329

Social Sciences	# of items
5. Psychology: BF 1-990	349
6. Anthropology: GN 1-696	145
7. Public Finance: HJ	137
8. Sociology: HM 1-221	153
Complete Listings	
1. Urban Planning. Urban Redevelopment: HT 166-177	550
2. Post Confederation Ontario History: F 5520-5547	317

* Total estimated population of English language, non-serial items N=16,910, with 8,601 (or 51%) in the Humanities, and 8,310 (or 49%) in the Social Science disciplines.

ANALYTICAL STUDY OF BOOK INDEXES ...UNDERWAY

From each monograph selected at the University of Toronto, we have xeroxed when available, indexes, table of contents and lists of illustrations, maps, figures, tables, etc. The preliminary analysis will consist of totals of the following:

1) actual number of monographs selected

2) number of monographs with an index

3) number of monographs without an index, table of contents, or lists

4) number of monographs that contain a table of contents but no index

5) number of monographs that have no table of contents or index but have some kind of list of illustrations, maps, or tables.

The above analyses will be broken down by L.C. class with totals for the entire sample.

In addition, an average number of index pages, table of contents pages, and lists of illustrations, maps, etc. will be computed for each class and the entire sample. We will also calculate the average number of entries per index page, table of contents page and list page for each L.C. class and the entire sample.

If possible, we will try to assess differences by date of publication. Although the sample was selected from post-1966 acquisitions, it may contain older items.

ANYONE INTERESTED?

By this time in your reading of our newsletter, you are probably wondering what difference our project will make to you in your library. In preparation for our answer to that question, we hope to sample check the holdings of other libraries against the data base we are creating from the University of Toronto collection. We suspect that you have many of these titles in your library and that you would be interested in searching such a data base on-line to determine how much more access it will provide to your collection than the access you presently have.

If interested in more details about this, drop us a line.

A NOTE OF APPRECIATION

Many people need to be thanked for encouraging us to embark on this project. We are very appreciative of your enthusiastic support. Without it, we would not have embarked on such a tedious but important task.

We hope that this and subsequent newsletters will keep you informed of our progress.

We also hope you will continue to be interested in our work and send us comments when you have ideas to share with us.